THE IMPUTATION OF COMPOSITIONAL DATA

David Judkins[1], Katie A. Hubbell, and Amy M. England, Westat, Inc.
Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Abstract

Compositional data come up in many settings in survey research and pose special problems for imputation. One prime example is expenditure data. The surveyor wishes to know how much was paid for a certain product or service and what financial resources were used to obliterate that debt. The special problems arise out of several features. First, all the potential sources must make nonnegative contributions. Second, the sum of the contributions must equal the charge. Third, the variety of missing patterns is astounding and definitely not nested or ignorable. Medicaid patients know very well that they paid nothing for a particular service, but they have no idea what the total charge was or who really did pay it (the state or the provider?). HMO patients are similar. On the other hand, there are people who know very well what the total charge is but don't know yet how much insurance will pay and how much they will have to pay themselves. We have developed a new algorithm, partially inspired by Gibbs Sampling. We describe the algorithm and present results from a small test dataset.

1

Introduction

The imputation of payment sources is a critical area for the Medicare Current Beneficiary Survey (MCBS) since the primary focus of the survey is on how Medicare beneficiaries meet the financial responsibilities for their medical care. From Medicare records, we know the total approved cost of covered services, and we know how much of that amount was paid by Medicare. However, it is the distribution of the balance across possible payment sources that is of primary interest. Furthermore, the total cost of uncovered services such as dental care and how those costs were met is also of high interest. Experience with the National Medical Care Expenditure Survey (NMES) indicates quite strongly that the missing data structure will not be nested.

Despite the importance of compositional data in financially-oriented consumption and business surveys, no systematic, general-purpose approaches have been developed for the imputation of missing compositional data. Although a fair amount of work has been done for establishment surveys at the Census Bureau, this work is largely unpublished and ad-hoc, relying heavily on the expertise of subject-matter specialists. With the large MCBS sample size and the low level of support for analytic staff at the contracting firm, this approach was not viable for MCBS.

The term "compositional data" appears to have been coined by chemists or geologists. Jointly, the two disciplines have published extensively on this type of data. However, their work typically assumes that each possible component of the mixture is present in at least minute quantities. (Some work does allow for a small probability that a component is actually totally missing, but the strategy is rather complicated.) Although this assumption may be reasonable in chemical and geological work, it is not reasonable for surveys about financial aspects of consumption. (For example, people without private health insurance are not going to receive any reimbursement from such insurance.)

Nonetheless, we did pick up some ideas from this field and combined them with ideas from Gibbs Sampling and from traditional hot-deck imputation methods to develop some new approaches to the problem. We developed three different approaches. However, due to space limitations, we only describe in detail the approach that we believe to hold the most promise.

In the following sections, we discuss notation, the algorithm and a couple of alternative ideas, development of an artificial database, results, and recommendations.

Notation:

$\zeta=(\delta,Y,Y_+)$ is the vector to be imputed, where

$\delta=(\delta_1,...,\delta_s)$ where $\delta_i=1$ if the i-th component is known to be present in a composition, $\delta_i=0$ if the i-th component is known to be absent from a composition.

$Y=(Y_1,..., Y_s)$ where Y_i is the level of the i-th component in the mixture, and

Y_+ is the total quantity of the mixture, measured in the same units as all the Y_i.

[1]The authors are all employed at Westat, Inc., Rockville, MD. The work was supported by the Office of the Actuary in the Health Care Financing Administration.

To aid in the imputation, the analyst will typically have a set of background variables available which provide predictive information about the composition. Let X be the matrix of values associated with such a set of predictor variables.

It is possible for any or all parts of ζ to be missing. Let $g=(g_1,...,g_s,g_+)$ where $g_i=1$ if Y_i is observed and 0 otherwise. Furthermore, let $h=(h_1,...,h_s)$ where $h_i=1$ if δ_i is observed (either clearly 0 or 1) and 0 otherwise. Let Ω_h be the set of distinct values of h realized in the sample. Let h^* be that element of Ω_h for which all the $h_i=1$; i.e., h^* represents perfectly observed δ.

The unique feature of compositional data that makes them so difficult to impute is that they must obey two constraints:

$$0 \quad Y_i \quad Y_+ \text{ for every i and} \tag{1}$$

$$\Sigma_i Y_i = Y_+. \tag{2}$$

The Skeleton of the Algorithm

The algorithm has an iterative aspect that was inspired by Gibbs Sampling. However, it is not a strict application of that technique.

The first step is to make sure that the reported data obey the constraints and that nothing can be filled in by simple subtraction or addition. Besides checking constraints 1 and 2, it is necessary to check that $Y_i>0$ implies $\delta_i=1$ and $Y_i=0$ implies $\delta_i=0$.

The second step is to impute δ. For each element h of Ω_h, conduct a separate hot-deck run to impute the missing portion of δ, where the donors and missing cases are matched on X and on the observed components of δ. Draw the donors from those with pattern $h=h^*$. At this point, δ is complete.

The third step is to come up with an initial feasible solution without worrying about how good the solution is. An initial solution is one where Y and Y_+ are complete, obey the constraints, and are consistent with δ. The hope is that, due to the iterative nature of the procedure, the starting solution is not very important. We used two different methods to complete ζ depending upon g. If $g_+=0$ (i.e., Y_+ is missing), then we sequentially imputed all the Y_i such that $g_i=0$, where each imputation was a simple hot-deck with δ_i and X as conditioning variables. After completion of Y, we imputed Y_+ as the sum of the imputed and reported Y_i. If, on the other hand, $g_+=1$, then we counted up the number of missing Y_i thought to be positive as $m=\Sigma_i\delta_i(1-g_i)$ and set each of the positive missing $Y_i=(Y_+-Y_{R+})/m$, where $Y_{R+}=\Sigma_j\delta_j g_j Y_j$ is the sum of reported elements of Y.

The fourth step is to re-impute Y_1 for each case where Y_1 and Y_+ were both originally missing. This is done with a hot deck conditioned upon δ_1, the sum of the other components of Y, and on X. After Y_1 is re-imputed, its new value is added on to the sum of the other components to obtain a new value for Y_+. This step is repeated for each of the Y_i. The motivation for the step is to improve the pair-wise consistency of the individual Y_i with the total, Y_+.

The fifth step is to re-impute the division of Y_1+Y_2 between Y_1 and Y_2 for all cases where both Y_1 and Y_2 were originally missing but known to be positive. This is done with a hot deck conditioned on $\delta_1,\delta_2,Y_1+Y_2$, and X. The hot deck actually imputes $P_1=Y_1/(Y_1+Y_2)$. The program then computes appropriate new values of Y_1 and Y_2. This step is repeated for each possible pair of components of Y. The motivation for the step is to improve the pair-wise consistency of the components of Y.

The fourth and fifth steps are then iterated until adequate convergence has been obtained.

More Motivation and Details on the Algorithm

Step 2. The maximum number of elements in Ω_h is 2^s. If s is large, running a separate hot-deck for each element of Ω_h may be impractical. We tested an iterative procedure that involved fitting logistic models on pairs of components of δ, conditioning on preliminary imputed values for the rest of δ and X. However, we ran into severe problems with structural zeros. Another possibility that we did not develop if s is large is to develop an iterative pairwise hot-deck procedure for the components of δ.

An issue that we didn't study at length was the choice of matching priority for the predictor variables. Typically, X and the observed portion of δ yielded more information than could actually be utilized in the hot-deck. (The full cross-product of all predictor variables led to cells so fine that there were some with only missing values.)

Step 3. There are many different ways to develop initial feasible solutions. We did not establish invariance of the final result to the initial solution. In fact, we doubt that the procedure always converges to the same solution. This is an area that probably deserves more research.

An important question is whether it is even necessary to resort to an iterative algorithm. We did develop and test a non-iterative algorithm, the core of which was an Aitchinson-type model for Y. The quality of the results on the simulated data set was excellent with this alternative. Unfortunately, the number of models that must be formed with this method rises with s even more sharply than does the number of hot-deck runs in step 2. Although computer time would be smaller with this noniterative approach, the amount of analyst time required to fit so many models is a fatal flaw.

Steps 4 and 5. We developed parametric alternatives to the hot-decks in these steps. Unfortunately, these parametric models required substantially more human intervention than the nonparametric hot-decks -- without improving the quality. In fact, the parametric results tended to be worse for the simulated data set that we tested.

As in step 2, we didn't pay much attention to the priority given to the predictor variables in the hot-deck matching. In step 4, Y_+-Y_i was categorized into quartiles. The categorized variable was giving the highest match priority after δ_i. A similar procedure was used for Y_i+Y_j in step 5.

Testing

We created an artificial dataset for use in development and testing of the three methods. The dataset was deliberately created to be plausible for the health insurance arena.

The first step was to create background variables that are assumed to be complete at the beginning of imputation. We used four binary variables and one continuous variable.

X_1 (INSURED) indicates whether case has health insurance

X_2 (COMPANY) indicates coverage by one of two imaginary health insurance companies

X_3 (SERVICE) indicates which of two services was performed on the case

X_4 (METRO) indicates whether the case lives in a metropolitan area.

X_5 is the income of the case. It was generated so that $\ln(X_5)$~N(9,2.25).

The next step was to generate Y_+, the total cost of the service. We used the model:

$$Y_{+i} \sim N(50+250X_{3i}-10X_{4i},\ 3+5X_{3i}+X_{4i}).$$

Note the lack of homoscedasticity. It is assumed that the variance in the cost is greater for one service than for the other and that the variance is also larger in metropolitan areas than in nonmetropolitan areas. Heteroscedasticity like this is likely to be present in MCBS and very difficult to detect.

Before creating the contributions to the bill from each source, we first created a "true" probe status for two of the three sources. For out-of-pocket expense, the probe was stochastically created such that

$$\text{logit}\{\Pr\{\delta_{1i}=1\}\} = \begin{cases} 3-5\delta_{2i}X_{2i} \text{ if } (Y_{+i}/X_{5i}) \quad 100 \\ 3\delta_{2i}-5\delta_{2i}X_{2i} \text{ otherwise.} \end{cases}$$

Note this leads to a very complex model for δ_{1i} that would probably not be ascertained in practice. One idea behind the model is that the likelihood of out-of-pocket expense for extremely expensive services is higher if insurance is paying part of the bill (δ_{2i}=1). The other idea is that one of the companies is very unlikely to require copayment for one of the services.

The second probe, for insurance coverage, was also stochastically generated such that

$$\text{logit}\{\Pr\{\delta_{2i}=1\}\}= 2-30(1-X_{1i})+3X_{1i}X_{3i}.$$

The ideas here are that the uninsured are extremely unlikely to have their medical expenses covered by their insurance and that both insurance companies are more likely to pay for one service than for the other service.

For Y_{1i}, we assumed that this variable represented the cost paid out of pocket by the case. For one type of insurance, we assumed a straight 20% copay. For the other type, we assumed a flat \$10 copay plus a sliding 10% copay of the balance. For the uninsured, we assumed that they paid the whole amount up to some income-related limit. This is spelled out in the following formulae:

$$Y_{1i} = \begin{cases} Y_{+i}/5 & \text{if } \delta_{1i}=\delta_{2i}=X_{2i}=1 \\ 10+(Y_{+i}-10)/10 & \text{if } \delta_{1i}=\delta_{2i}=1 \text{ and } X_{2i}=0 \\ \min\{Y_{+i}, X_{5i}/100\} & \text{if } \delta_{1i}=1 \text{ but } \delta_{2i}=0, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

If the service was covered by insurance, then the amount paid by insurance was assumed to be the total bill, less the copay. Formulaically, this may be written as

$$Y_{2i} = \begin{cases} Y_{+i}-Y_{1i} & \text{if } \delta_{2i}=1, \\ 0 & \text{otherwise.} \end{cases}$$

Whatever was left of the bill was assumed to be picked up by other sources. These other sources could include public assistance, private assistance, forbearance by the provider, et cetera. Thus

$$Y_{3i} = Y_{+i}-Y_{1i}-Y_{2i} \text{ and}$$

$$\delta_{3i} = 1 \text{ if and only if } Y_{3i}>0.$$

Having created the "truth" the next step was to simulate nonresponse. Nonresponse was possible on any of seven dimensions. The first dimension was not being able to report the total bill. Here we used

$$\text{logit}\{\Pr\{g_{+i}=1\}\}=1 +3(1-X_{1i})-2Y_{+i}.$$

Note the dependence of the probability of nonresponse on the amount of the bill. This type of nonresponse is nonignorable and impossible to correct for unless the model is known in advance. We deliberately used such a model to make it tough on the methods. Given that toughness, we shouldn't expect any of the methods to work perfectly.

For not knowing the amount paid out of pocket, we used

$$\text{logit}\{\Pr\{g_{1i}=1\}\}= 1.5 + (1-X_{1i})/10 - \delta_{2i}/2 - (.4)X_{1i}X_{21}X_{31}.$$

Note the complex interaction term. That will make modeling difficult.

For not knowing amount paid by insurance, we used

$$\text{logit}\{\Pr\{g_{2i}=1\}\} = -2.5 +g_{i1} +3g_{+i} +30(1-X_{1i})- (1.5)X_{1i}X_{21}X_{31}$$

Again, there is a complex interaction term. Also note the relationship with nonresponse on amount paid out of pocket and nonresponse on amount of total bill. Finally, note that the large coefficient forces everyone without insurance to know that insurance didn't pay anything.

For not knowing the amount paid by other sources, we used

$$\text{logit}\{\Pr\{g_{3i}=1\}\}= -2 + g_{1i} +2g_{2i} +g_{+i} + +99[g_{1i}g_{2i}g_{+i} - g_{1i}g_{2i}(1-g_{+i}) -g_{1i}(1-g_{2i})g_{3i}-(1-g_{1i})g_{2i}g_{3i}].$$

The extremely large 99 forces response on the amount paid by other sources if all other amounts are known. It also forces nonresponse if exactly two of the three other amounts are missing.

Lastly, nonresponse sometimes occurs on the probes, as well, where the person doesn't even know if a particular source paid any part of the bill (most likely with proxy respondent). Here we used:

$$\text{logit}\{\Pr\{h_{1i}=1\}\} = 1+ 400g_{1i}.$$

$$\text{logit}\{\Pr\{h_{2i}=1\}\} = 1+ 400g_{2i} - 2(1-g_{1i}).$$

$$\text{logit}\{\Pr\{h_{3i}=1\}\} = 1+ 400g_{3i} - 2(1-g_{1i}) - 2(1-g_{2i}) - 4(1-g_{1i})(1-g_{2i})$$

The 400s force response on the probe if the amount is known. Other assumptions are that if a person doesn't know the amount paid out of pocket, then he/she is less likely to even know whether insurance paid any part of the bill.

Results

Table 1 summarizes the results. The parameter of interest is given in the first column. The next column gives the truth for the particular data set that was generated. Then missing rates are given. Note that missing rates are higher for correlations than for other statistics since both variables have to observed in order to compute the correlation.

The "Observed" column indicates what the naive analyst would obtain without any imputation given the induced nonresponse. Note that two of the means are gravely biased, as are one of the correlations, and two of the kurtosis factors.

The "Imputed" column gives the results for our preferred algorithm. In general, our method worked better than no imputation. Note in particular, the dramatic improvements in the means of Y_2 and Y_3, the correlation between Y_2 and Y_+, the standard deviation of Y_2, and the kurtosis of Y_2 and Y_3. There was no major increase in bias for any of the parameters due to imputation.

Further Study

Our only real remaining reservation about our algorithm is the amount of computer run time that it might require with large s and a large number of cases. For MCBS, we anticipate s=9 and several hundred thousand records. We plan to run a series of progressively larger tests to determine just how computer time the algorithm will require in a realistic situation.

Table 1. Evaluation of algorithm on 200 cases with s=3

Parameter	Truth	Missing rate	Observed	Imputed
Means				
δ_1	0.69	5%	0.69	0.70
δ_2	0.57	15%	0.50	0.56
δ_3	0.27	31%	0.33	0.26
Y_1	35.6	25%	38.6	36.0
Y_2	63.8	36%	37.4	63.2
Y_3	52.5	50%	70.4	52.8
Y_+	152	40%	165.8	151.9
Correlations:				
Y_1 and Y_2	-0.14	49%	-0.14	-0.13
Y_1 and Y_3	-0.11	54%	-0.18	-0.11
Y_2 and Y_3	-0.32	58%	-0.29	-0.32
Y_1 and Y_+	0.32	55%	0.34	0.33
Y_2 and Y_+	0.47	47%	0.24	0.46
Y_3 and Y_+	0.53	57%	0.63	0.53
Standard Deviations				
Y_1	65	25%	67	65
Y_2	100	36%	82	99
Y_3	105	50%	115	105
Y_+	124	40%	126	123
Skewness				

Y_1	3.2	25%	3.0	3.2
Y_2	1.5	36%	2.4	1.5
Y_3	1.7	50%	1.2	1.7
Y_+	0.3	40%	0.1	0.3
Kurtosis				
Y_1	9.7	25%	8.3	9.5
Y_2	0.6	36%	4.3	0.6
Y_3	1.0	50%	-0.4	0.9
Y_+	-1.9	40%	-2.0	-1